MATCHING ECONOMIC EFFICIENCY AND ENVIRONMENTAL SUSTAINABILITY: THE POTENTIAL OF EXCHANGING EXCESS CAPACITY IN CLOUD SERVICE ENVIRONMENTS

Completed Research Paper

Christoph Dorsch
Augsburg University
FIM Research Center
Universitätsstr. 12, 86159 Augsburg
Germany
christoph.dorsch@wiwi.uni-augsburg.de

Björn Häckel
Augsburg University
FIM Research Center
Universitätsstr. 12, 86159 Augsburg
Germany
bjoern.haeckel@wiwi.uni-augsburg.de

Abstract

Excess capacity is a major problem for service providers. While manufacturers e.g. can produce on stock to fully utilize their capacity, the service industry traditionally faces the problem of idle capacity resulting in economic and environmental inefficiencies. Recent technological developments offering dynamic information and integration capabilities may help, as they enable an on-demand exchange of excess capacity. To examine the possible benefits of corresponding excess capacity markets, we examine a capacity related optimization problem of a service provider with and without relying on excess capacity. Therefore we build a mathematical model based on queuing theory which is evaluated with a discrete-event simulation applying the situation of providers for banking transaction services. By solving the optimization problem we found reasonable benefits of excess capacity markets concerning the economic and environmental perspective. Being a first quantitative approach, the model thereby builds the basis for empirical validation and further theoretical research.

Keywords: IT-enabled organizational capabilities, Service-Oriented Architecture, Design Science, Inter-organizational systems, Service supply chain, Sustainability, Volatility

Introduction

The ex ante planning of capacity is a widely recognized challenge in different streams of economic literature. Determining an appropriate level of capacity ex ante thereby is a crucial task, especially in case of highly volatile demand, non-adjustable capacity in short-term and a time critical execution due to customer needs (Adenso-Diaz et al. 2002). In production management, a common strategy to cope with this challenge is building up excess inventories in time frames of low demand and reducing these inventories in time frames of excess demand. However, the ex ante planning of capacity shows up even more difficult for service providers, as services are not storable in general (Chesbrough et al. 2006; Rai et al. 2006). Thus, in contrast to physical goods the building of excess inventories is not a possible strategy for services providers. At the same time, service providers usually have to guarantee service level agreements (SLA) to their customers (e. g. a maximum execution time for each incoming order) and therefore are confronted with contractual penalty payments in case of violating the committed SLA. SLAs are especially common when orders are time critical as this is the case e. g. for banking transaction services. Against this background, seeking to optimize the level of assigned capacity for offered services, a service provider is faced with a trade-off (Bassamboo et al. 2010a; Bassamboo et al. 2010b): Assigning a high level of capacity allows the buffering of temporarily peaks in customer demand but may result in idle costs in time frames of low demand. Assigning less capacity avoids idle costs but may result in waiting costs due to contractual penalties in time frames of high demand. Consequently, from an economic point of view, a service provider aims at minimizing the total processing costs (consisting of idle and waiting costs) by choosing an appropriate level of capacity (Bassamboo et al. 2010a). As this trade-off is hard to solve, service providers tend to hold significant overcapacities to cover peak demand and with that avoid high contractual penalties (Liu et al. 2010). The resulting underutilization of esp. IT capacity is highlighted by the fact that in typical data centers, resource utilization is only 20-30% on average (Bohrer et al. 2002). At the same time, industry leaders estimate nearly $450 billion US-$ are being spent annually on new data center space (Cook 2012). Even though resource utilization can be increased, e. g. through virtualization leading to more scalable IT infrastructures (Vykoukal 2010), unutilized IT resources stay at an unsatisfying high level.

Considering these key figures, the tremendous environmental effects of unutilized *IT capacity* become obvious. In addition, considering capacity planning for services, not only IT capacity is of relevance. Usually also *personnel resources* have to be taken into account as even mainly IT-driven services often require manual interventions. Since personnel resources are even less scalable in short-term compared to IT resources, the described trade-off within capacity planning is further exacerbated for services with manual interventions. Consequently, the peculiarities of services involve tough challenges within capacity planning of service providers often resulting in an inefficient use of personnel as well as IT resources. However, enabled by technological developments new possibilities arise that may help to overcome the described problems. With the growing diffusion of service-oriented infrastructures suitable for the integration of web services as well as corresponding description languages (e. g. WSDL) and standards for data exchange (e. g. XML, EDIFACT), a dynamic integration of business partners has become considerably easier (Grefen et al. 2006; Moitra et al. 2005). Dynamic integration capabilities thereby provide the basis for establishing partner relationships on-the-fly in an adaptive, fine-grained way. In this way, the idea of a dynamic business process outsourcing as well as the closely related *Business Process as a Service (BPaaS)* are meanwhile well established concepts, meaning that partner relationships can change frequently supported by automated integration mechanisms. Especially for highly standardized IT-driven services even the on-demand integration of external service providers meanwhile is a feasible alternative (Grefen et al. 2006) with corresponding economic potential as already demonstrated in Dorsch et al. (2012). Based on these economic potentials, the technological developments lead to the establishment of service marketplaces where firms that offer or/and demand certain services can interact in a highly dynamic manner (Grefen et al. 2006).

With respect to the capacity optimization problem focused in this paper, such IT-driven marketplaces may offer substantial benefits as underutilized IT and personnel capacity can be exchanged: Cloud service providers can continuously provide relevant information about their excess capacity which can be evaluated by other service providers automatically. Based on the provided information other cloud service providers then can decide whether to use available excess capacity. The benefits of an excess capacity market are twofold, offering *economic as well as environmental benefits*: First, for a cloud service

provider the opportunity arises to route excessive demand to external service providers with available excess capacity. This opportunity therefore might help to mitigate the described trade-off between idle costs and waiting costs, consequently leading to a reduction of total processing costs (*economic efficiency*). Second, the opportunity to exchange excess capacity on short notice allows for a more sustainable usage of capacity regarding the whole cloud service provider market. In particular, a cloud service provider is supposed to reduce his in-house capacity, as he can route excessive demand to the excess capacity market in peak times. Consequently, unutilized overcapacities are supposed to be reduced for a single cloud service provider as well as for the whole cloud service provider market (*environmental sustainability*). However, along with these potential benefits, the option of using excess capacity also bears additional risk compared to in-house capacity: As only excess capacities are offered that otherwise remain idle, usually no SLA is committed. Consequently, a service provider demanding non SLA backed excess capacity on short-notice faces the risk of only getting served as soon as capacity is available on the market. That might cause delays for external routed demand and thus waiting costs. Hence, when evaluating the potentials of using excess capacity, this additional risk has to be considered.

The corresponding effects have not gained much attention in literature so far. Hence, in the paper at hand we will analyze the potentials of exchanging excess capacity in a cloud service environment. In doing so, we will refer to the capacity planning of a cloud service provider and develop an optimization model based on queuing systems. Within our optimization model we will in particular take into account the effects of an excess capacity market on economic efficiency as well as on environmental sustainability. By analyzing our optimization model, we focus on the following research questions:

How does the opportunity to use excess capacity in a cloud service environment affect the capacity planning of a cloud service provider?

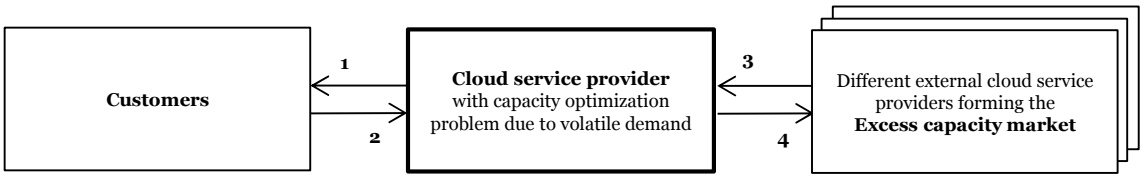
To what extent does the existence of an excess capacity market in particular help to improve economic efficiency as well as environmental sustainability?

The remainder of this paper is organized as follows: First, we describe the general setting of our optimization problem and provide a review of the related literature as well as an overview of the research methodology. Afterwards we develop a formal model of this optimization problem as the basis for further analysis and introduce a possible application scenario adapting the situation of a provider for banking transaction services. By executing a discrete-event simulation we are then able to analyze and present the effects of exchanging excess capacity on economic efficiency and environmental sustainability. Finally, we discuss limitations and directions for subsequent research and summarize the findings of the paper.

General Setting, Related Work and Research Methodology

Before introducing our optimization model, we will first describe the general model setting in more detail. Second, we will discuss related literature and outline the research methodology our paper is based on.

General Setting and Preconditions for an Excess Capacity Market



1 The **cloud service provider** offers an **IT-driven service** *with* committed service level agreement (SLA) to the **customers**.

2 The **customers** use this **service** from the **cloud service provider** sending orders characterized by *volatile demand*.

3 The **excess capacity market** offers **excess capacity** *without* committed SLA to the **cloud service provider**.

4 The **cloud service provider** uses excess capacity if the **excess capacity market** *enables "preferable execution"*.

Figure 1. Cloud Service Relationship with Excess Capacity Market

We consider a cloud service provider that offers an IT-driven service to his customers. As the execution of the service is time critical for the customers, the cloud service provider commits a SLA (as indicated by arrow number 1). The offered service thereby is characterized by a volatile customer demand (arrow number 2). Some of the activities of the service are executed by the IT system of the cloud service provider, whereas others require manual interventions. As neither the IT capacity nor the personnel resources are fully scalable, meaning that they cannot be adjusted in short-term, the cloud service provider faces a capacity optimization problem for his in-house operating unit. That is because the cloud service provider wants to avoid costly violations of the committed SLA caused by capacity shortages in times of peak demand on the one hand and idle costs in times of low demand on the other hand. In addition to draw on in-house capacity only, the cloud service provider can use available excess capacities from different external cloud service providers that form the so called excess capacity market (arrow number 3). Hence, an additional execution path arises for the cloud service provider, as especially peak demand can be routed to external cloud service providers. As excess capacity usually is not SLA backed, this option tends to be more risky (due to possible delays) but cheaper in contrast to in-house capacity (as it is excess capacity otherwise remains idle). Thus, to decide whether an external execution is preferable compared to in-house execution, the cloud service provider for each incoming order has to weight the costs for external execution as well as the risk of possible delays against the total processing costs of the in-house unit (arrow number 4).

To operationalize the use of excess capacity, some preconditions regarding the supply of information have to be fulfilled. In particular, a service provider that intends to use excess capacity has to determine which external providers have sufficient excess capacity available at the very moment or by when capacity will be available respectively. Thus, the service provider's IT-platform has to allow a continuous, mostly automated evaluation of external providers and all relevant information has to be provided by the external providers' market. As already outlined in the introduction, this is enabled by technological developments that foster strong on-demand integration capabilities. The provision of information is e. g. supported by high-level frameworks for information exchange like ebXML and RosettaNet as well as by various B2B gateways that are offered by product vendors like e. g. Oracle and IBM. In recent years, in particular the web service paradigm coming along with service repositories and well described services based on standardized description languages have evolved as one of the primary standards for a dynamic evaluation and integration of service providers (Grefen et al. 2006). Based on these technological developments, on the one hand a decentralized information exchange between various service providers regarding the usage of excess capacity is enabled. On the other hand, a more centralized approach is enabled by the development of (on-demand) service marketplaces like e. g. SAP Service Marketplace, HubSpot or Zimory where firms that offer or/and demand certain services can interact in a highly dynamic manner (Grefen et al. 2006, Weinhardt et al. 2009). As these service marketplaces enable a coordinated interplay of customers and providers they can also be used to foster an efficient usage of capacity on the external providers' market by dynamically matching (excess) capacity demand and supply. A dynamic matching thereby can be supported by dynamic pricing mechanisms like e. g. auctions that are widely discussed in literature (Anandasivam et al. 2009, Weinhardt et al. 2009, Wurman 2001).

Related Work

The problem of capacity planning under uncertain demand for non-storable goods and in particular services has already been addressed in several papers. Especially there is a broad literature stream focused on the topic of call center outsourcing which reflects a common example for capacity planning for services. These papers usually distinguish between two basic sourcing models a company can built its capacity planning on, namely the use of volume-based contracts or capacity-based contracts (e. g. Gans et al. 2007, Aksin et al. 2008). Thereby volume-based contracts ("pay-for-job") involve payments only for capacity that is used, whereas capacity-based contracts ("pay-for-capacity") involve payments for capacity whether it is used or not. In terms of our model setting, capacity-based contracts correspond to the option of using in-house capacity as the cloud service provider has to pay fixed costs for in-house capacity whether it is utilized or not. On the other hand volume-based contracts in general correspond to the use of excess capacity from the market, as the cloud service provider pays for each order routed externally. Aksin et al. (2008) consider a call center outsourcing relationship where a service provider can choice between a volume-based and a capacity-based contract offered by a contractor that aims at determining the optimal capacity levels. The paper determines optimal capacity levels and partially characterizes optimal pricing

conditions under each contract. The paper of Gans et al. (2007) also distinguishes between volume or capacity based outsourcing contracts and analyzes the centralized capacity and queuing control problem. Further papers dealing with outsourcing decisions in a service setting are e. g. Cachon et al. (2002), Allon et al. (2006) and Ren et al. (2008). Cachon et al. (2002) study the competition between two service providers with price- and time-sensitive demand by modeling this setting as a queuing game. One of their core results is that scale economies provide a strong motivation for outsourcing. The work of Allon et al. (2006) analyzes the situation of retailers who are locked in price and waiting-time competition and have the option to outsource their call center service to a vendor. Thereby, among others volume-based contracts and their effects on supply chain coordination are analyzed. Ren et al. (2008) study contracting issues between a client and a vendor (call center) that does outsourcing work for the client. Within their paper Ren et al. (2008) analyze contracts the client can use to induce the call center to choose staffing and effort levels that are optimal for the supply chain. Our approach differs from the literature outlined above as these papers consider volume-based contracts that are backed by a SLA. In our approach we take explicitly into account the usage of excess capacity that is usually not SLA backed and thus tends to be cheaper but more risky on the other hand.

Other papers closely related to our idea of an excess capacity market consider so called surplus markets in the areas of production and supply chain management. Among others, Dong et al. (2005) study IT-driven markets for surplus components, which allow manufacturers with excess component inventory to sell to firms with a shortage. They derive conditions on demand uncertainty that determine whether a surplus market will increase or decrease supplier profits. Another paper dealing with flexibility of supplier markets is Lee et al. (2002). The paper investigates the impacts of a secondary market, where resellers can buy and sell excess inventories. For that, Lee et al. (2002) derive optimal decisions for the resellers regarding their ordering policies and analyze the effects of the secondary market both on the sales of the manufacturer and the supply chain performance. These papers are closely related to our approach regarding the basic idea of a surplus market. As a fundamental difference to our approach these papers are concerned with physical products and thus are more concerned with the possible trading of physical excess inventories and its implications on capacity planning. Thereby a time critical delivery time of products is not considered. However, in our approach we focus on the capacity planning problem of a cloud service provider, who provides a non-storable and time-critical service to his customers.

Furthermore, in our approach we do not only consider the economic effects of an IT-driven excess capacity market but also its potentials regarding environmental sustainability. The analysis of environmental effects of IS in general is a comparatively new field of research. Recent works in this area are e. g. provided by Chen et al. (2009), Watson et al. (2009), Watson et al. (2010) and Melville (2010). Watson et al. (2010) labeled this field of research as energy informatics. In their paper they call for Green IS initiatives and demand the IS community to fulfill their social responsibility which has long been neglected. In addition, Melville (2010) and Watson et al. (2009) claim business organizations to use their economic power to promote climate change and environmental sustainability by making use the transformative power of IS. Chen et al. (2009) describe that besides other factors in particular financial concerns influence an organization's decision to adopt green IS. However, a quantitative approach to capture the environmental effects of IS is not presented in those papers. Rather, there are only a few papers that especially address the topic of so called "green cloud computing" (Baliga et al. 2010, Singh et al. 2009, Liu et al. 2009, Vereecken et al. 2008). These papers focus on IT-driven environmental effects that either stem from the energy consumption of data centers that host cloud computing services or the energy consumption of transmission and switching networks within cloud environments. In contrast to these papers, in our approach we do not only consider the classical "IT Cloud" but instead focus on the so called "Service Cloud" that also comprises the human part of processing work and service delivery and that is manifested in concepts like BPaaS. In particular, we focus on a more efficient usage of IT and personnel capacities *enabled by new developments in IT* and not solely on the efficient usage of IT capacity as investigated by the named papers. Consequently, we aim to capture a broader view on the environmental effects of cloud service environments by considering both, IT and personnel capacities.

Summarizing, to our best knowledge an integrated analysis on both the economic and the environmental effects of an IT-enabled excess capacity market in a cloud service environment has not been addressed in literature so far. Therefore, we aim on contributing to the closure of this research gap by developing an optimization approach that allows for analyzing both effects.

Research Methodology

To answer our research questions, we apply a typical design-science driven research approach. The basic paradigm of design-science research is to solve organizational problems and to gain knowledge of a problem domain by creating and applying specific artifacts (Hevner et al. 2004, Peffers et al. 2007). Thus, strongly related to the design-science research guidelines outlined in Hevner et al. (2004) we start in the following with developing an artifact, which in our case is an optimization model to quantify the economic and environmental effects of using excess capacity. As outlined in Hevner et al. (2004), building a mathematical model is one common way to represent an artifact in a structured and formalized way. In a second step we evaluate our optimization model by combining an experimental and a descriptive design evaluation method which is a widely used approach for evaluating artifacts based on mathematical models (e. g. Wacker 1998): We describe a detailed real world scenario, in which a provider for banking transaction services is confronted with a capacity optimization problem (descriptive design evaluation method). Based on that scenario we then perform a discrete event simulation study (experimental design evaluation method). As the major goal of design-science research is utility (Hevner et al. 2004), within our evaluation approach we aim to show how our optimization model can be applied to a specific scenario and how the economic and environmental effects of using excess capacity can be valuated based on this model. Such a valuation might provide the basis for investment decision regarding the creation of on-demand integration capabilities. By developing and evaluating a specific artifact we therefore provide a thorough design-science research process in the sense of Hevner et al. (2004). However, to further contribute to the complementary research cycle between design-science and behavioral-science further research regarding the evaluation of our model in a given organizational context might be insightful (Hevner et al. 2004, McKay et al. 2012). In the course of that, in particular various empirical evaluation methods like e. g. case studies, field studies or field experiments as well as statistical sampling could be applied (Hevner et al. 2004, Wacker 1998). The adopted approach in this paper is also closely related to the basic idea of the research cycle of Meredith et al. (1989) who emphasize that describing non-examined research areas qualitatively and mathematically and thus predicting first results provides the basis for generating hypothesis that can be tested within future empirical research. Hence, to address this point and to outline next steps for further research, we will explicitly discuss various research directions regarding our optimization approach that might be addressed by applying additional evaluation methods at the end of the paper.

Modeling the Cloud Service Relationship and Excess Capacity Market

The following model forms the basis to examine the effect of *exchanging excess capacity on economic efficiency and environmental sustainability* within the setting illustrated in Figure 1. Starting point is the underlying economic capacity optimization problem of the cloud service provider resulting from the volatile arrival rates of time-critical orders. Then we describe functionality and interaction of the in-house operating unit and the excess capacity market as well as their parameters relevant to state the optimization problem. Finally, we introduce a routing algorithm necessary to analyze the model which is done in the following section with an application scenario based on a real world example.

Underlying Capacity Optimization Problem

The *cloud service provider* offers an IT driven service to its customers. This service is represented by a business process with different activities which are executed on orders the *customers* submit into the IT system of the cloud service provider. Some activities are executed *automatically within an IT system* while *other activities need manual interventions* executed by employees. Each incoming order triggers the execution of all activities. After the last activity is finished, the order is returned to the customer. The time frame between accepting and returning the order is called *processing time*.

A service level s (e. g. a maximum processing time with monetary compensation for each time unit the order exceeds this limit) is guaranteed to the customers regarding this processing time. Any order which does not keep up to this service level causes costs subsumed within c_s. The arrival rate λ, i. e. the number of time-critical orders sent from the customers per unit time, is random. This might lead to a bottleneck because all activities of the service are executed by an *in-house operating unit* with limited execution capacity which cannot be adjusted on short notice following the volatile demand. Based on historical data

and contractual agreements respectively the statistical distribution of λ can be approximated. The planning horizon considered is finite and divided into equidistant time units.

Taking these characteristics into account, the cloud service provider has to decide ex ante about the capacity (i. e. the number of orders $y \in IN_0$ which can be handled simultaneously) it allocates to the in-house operating unit, which minimizes the *total processing costs c* for the service. The simplified objective function for this discrete optimization problem therefore reads

$$\min_{y} c(\lambda, y, s)$$

The cloud service provider faces the following basic trade-off: Allocating too much capacity to the in-house operating unit causes excessive costs of (idle) capacity. Providing too little capacity result in long waiting times causing excessive follow-up costs regarding the service level guaranteed to the customers.

With regard to the main characteristics (i. e. random demand, limited capacity etc.), it is appropriate to model the capacity optimization problem using queuing theory. The in-house operating unit of the cloud service provider then can be regarded as a queuing system executing incoming orders. In doing so, we follow other authors, e. g. Braunwarth et al. (2010), Chen et al. (2010) or Hlupic et al. (1998) who relied on queuing theory to model and examine similar systems. Furthermore, using queuing theory a wide range of possible settings of the underlying service (e. g. limited business hours, day and night operation, overtime work, alternating order execution times) can be easily considered. In the following, we therefore rely on the basic assumptions of queuing theory as e. g. described in Gross et al. (2008), extending them by parameters and functions to specify the setting illustrated in Figure 1.

In-house Operating Unit and Excess Capacity Market

The execution of an order starts immediately with its arrival in the cloud service provider's IT system unless all units of capacity within the in-house operating unit are busy. Otherwise each incoming order lines up in an infinite waiting queue. The queued orders will be executed immediately after free capacity is available according to the first in/first out principle. One order uses at least one unit of capacity for this time frame. Free units of capacity are idle or can be used to accelerate the execution of orders by assigning more than one unit of capacity to an order (networking effects). The time frame the order stays in the queue in front of the in-house operating unit is called *waiting time*. The time frame between the beginning of the first activity and the end of the last activity of the service is called *execution time*. Waiting and execution time sum up to the *processing time* mentioned above. Hence, long waiting times might lead to processing times which do not keep up to the service level agreed and cause corresponding costs.

In addition to the in-house operating unit, there are different *external cloud service providers* which are able to execute the activities of the business process, too. These external providers offer their excess capacity (IT capacity as well as employees) for temporary use forming a virtual *excess capacity market*. This excess capacity can be used by the cloud service provider to execute e. g. peak loads otherwise would wait for execution in the queue in front of the in-house operating-unit.

On this market, capacity cannot be booked in advance as only excess capacity is offered. Also, no service level guarantees a constant availability of capacity. Rather, the external cloud service providers continuously announce the time frames until free units of excess capacity will be available. Based on this information the cloud service provider then decides whether he concludes a temporary contractual agreement which guarantees the execution of an order within the announced time frame. The availability of capacity on the excess capacity market therefore changes over time and there is a non-negligible and risky waiting time to be considered when relying on the market.

Modeling the excess capacity market as a second queuing system to execute incoming orders, the in-house operating unit in combination with the excess capacity market forms a *service system* offering two execution paths for incoming orders (see Figure 2 on the next page). Thereby the cloud service provider now has the choice whether it routes an incoming order to the in-house unit or the excess capacity market. This decision has to take place during operations for every single order immediately at the time it arrives. The provider's IT system then has to evaluate automatically which *execution path* offers a "preferable" execution at that very moment and has to route the order to the respective path. As we are optimizing total processing costs, a "preferable execution" is given, if the corresponding costs of the

respective path are lower than for the other path (see subsection "Routing Algorithm Enabling an Automated Routing Decision" for details).

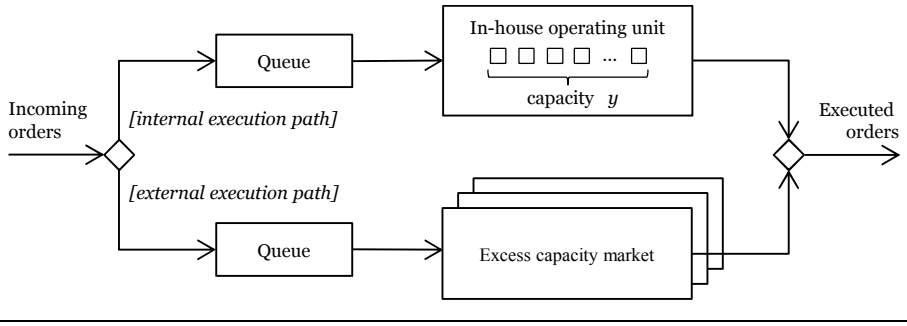


Figure 2. Service system with two queuing systems (in-house operating unit and excess capacity market)

To enable this automated routing decision, two requirements must be fulfilled: First, the excess capacity market (i. e. every single external cloud service provider) has to provide *continuously* all relevant information for the decision whether to use excess capacity or not (e. g. current waiting times until an unit of capacity will be available for temporary use, current costs for capacity usage etc.) which has to be evaluated *automatically* by the cloud service provider's IT system. Second, a *quick and frictionless* integration between the IT systems of the cloud service provider and the external cloud service providers must be possible, when using excess capacity on short notice. This is where the enabling technologies and standards connected to architectural concepts like service-orientation and web-services (e. g. service repositories and well described services based on standardized description languages) become relevant and we assume that suitable technologies as described in the introduction are established and running.

As one of the most important information, the exogenous waiting time for capacity on the excess capacity market has to be provided from all external cloud service providers. The time frame a denotes the time, an order has to wait in the queue in front of the excess capacity market until the next unit of excess capacity is offered to execute the activities of the service. With $a > 0$ orders cannot be executed immediately and the exogenous waiting time might be too long to keep up with the service level agreed to the customers causing corresponding costs.

Total Processing Costs and Detailed Objective Function

To determine the total processing costs required by the objective function, some additional parameters are necessary specifying the in-house operating unit and the excess capacity market:

The execution time t_i of the in-house operating unit for one order depends on its individual characteristics. Based on historical data, the statistical distribution of t_i is stated. There are fixed costs c_f per unit capacity but no additional variable costs. The total number of orders finally executed in-house is denoted with o_i.

The fixed costs considered for one unit of capacity contain recurring costs of capacity e. g. wages of employees, running costs for the IT system and other equipment, overhead costs as well as all non-recurring initial costs building up this capacity.

The execution time t_e of the excess capacity market for one order depends on its individual characteristics. Based on historical data the statistical distribution of t_e is stated. There are no fixed costs but variable costs c_e which come up using excess capacity. As prices are endogenous and may change during operations, the respective price c_e has to be provided along with the information about the waiting time a as described above. The total number of externally routed orders is denoted with o_e.

In doing so, we assume a similar execution time for all external cloud service providers. As we are focusing on standardized services and the individual characteristics of each order are already considered,

this seems acceptable. The variable costs include not only the price for order execution to be paid to an external cloud service provider, but also the costs related with the evaluation of the market and integration of the external cloud service provider.

Considering the underlying capacity optimization problem and all additional characteristics described above, we are now able to determine the total processing costs the cloud service provider faces and the detailed objective function reads

$$\min_{y} c = c_f\, y + c_e\, o_e + c_s(\lambda, y, o_i, o_e, s, t_i, t_e, a)$$

Hence, the total processing costs considered in this model consists of the fixed costs of in-house capacity, the variable cost of using excess capacity and the costs resulting from orders which could not keep up to the service level agreement.

The optimization problem is still related to the amount of capacity, the cloud service provider allocates to the in-house operating unit. Adding the excess capacity market only changes the total processing costs which now have to consider an additional trade-off for the cloud service provider: During operation it has to balance between the different processing costs for an incoming order when using the internal or the external execution path. Solving the objective function for integer values of y result in the optimal amount of capacity, the cloud service provider should allocate in-house to minimize the total operating costs.

Routing Algorithm Enabling an Automated Routing Decision

To enable the automated routing decision and thereby to solve the optimization problem it is not sufficient to evaluate the two queuing systems representing the in-house operating unit and the excess capacity market separately. Rather, the service system has to be evaluated as a whole as the two queuing systems interact during operations influencing the waiting times.

Although queuing theory provides a strong mathematical foundation, this cannot be done analytically since the two queuing systems have different characteristics, especially concerning their distribution of processing times. They cannot be integrated to a service system for which a mathematical model offers an analytical solution. This also explains why the costs c_s resulting from the service level agreed to the customers could not be specified in a closed term within the detailed objective function.

In fact, the automated routing decision requires an order routing algorithm which links the two interacting queuing systems and decides about the execution path for an incoming order. With regard to the cost-based optimization problem, this decision is made based on processing costs. With each arrival of an order the processing costs of both queuing systems are evaluated and the path with lower processing costs ("preferable execution") is chosen. The routing algorithm therefore works as follows: The algorithm first determines the processing time for each queuing system. For the in-house operating unit this is easily determinable as the state of the system is known: It depends on its capacity, the arrival rate of orders and the execution time. For the excess capacity market, the time frame a until free capacity will be available, has to be determined. Second, if this processing times result in a violation of the service level agreed, networking effects have to be calculated. The trade-off between higher execution costs and a possible reduction of penalties from the service level agreement build the basis for this decision. Third, having determined processing times and considered possible supporting networking effects, the processing costs of each execution path for the respective order can be determined.

For further analysis and validation of our model, we now introduce a possible application scenario based on a real world example and perform a discrete-event simulation. It implements the model setting described above and the necessary routing algorithm to derive interpretable results. A discrete-event simulation thereby is an established method analyzing queuing systems (Gross et al. 2008). Within this simulation the interaction of both queuing systems can be evaluated and a simulation based optimum (referred to as "optimal capacity" hereinafter) can be determined. The effect of exchanging excess capacity on economic efficiency and environmental sustainability then can be examined if the capacity optimization problem is solved with and without the excess capacity market.

Model Analysis: The Payment Transactions Process

To analyze the setting of a cloud service relationship with excess capacity market and to reveal the potential of exchanging excess capacity on economic efficiency and environmental sustainability we implement the model described above with an application scenario adapting the situation of a provider for banking transaction services. The business process considered is the payment transactions process. It includes all necessary activities to execute payment orders like bank transfers, direct debits, checks, drafts and returns as well as debit and credit card payments as illustrated simplified in Figure 3 (shaded part).

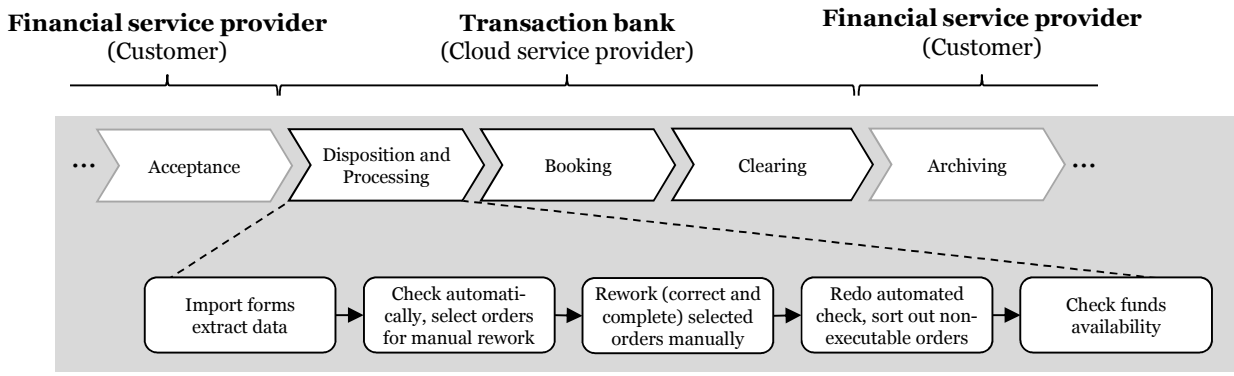


Figure 3. Cloud Service Relationship of the Payment Transactions Process

Cloud Service Relationship of the Payment Transactions Process

This process is a typical application scenario addressed with our model. It is an *IT driven service* most financial service providers meanwhile *source as a service* from a specialized business partner called "transaction bank". A large number of orders *characterized by volatile demand* have to be processed in time to meet regulatory standards and to avoid losses of interest. E. g. one of Germany's market leaders in payment transaction processing with a market share of about 20 % processes an average of 30 million transactions a day. The corresponding volume of money transferred is about EUR 120 billion. Therefore *detailed service levels* concerning the time frame for execution are agreed between the financial service provider and the transaction bank. With few exceptions the transaction process is fully digitalized and it is highly standardized through regulations, cross-company agreements and in Europe most recently through the introduction of the Single European Payment Area. The corresponding cloud service relationship between the financial service provider and the transaction bank is outlined in Figure 3 (non-shaded part).

Allocating IT capacity and employees to the in-house operating unit charged with processing the payment orders is an important optimization problem for the transaction bank. As the margins are small, the capacity of the in-house operating unit should be kept as small as possible to reduce the corresponding costs to a minimum. However, the limited time for executing the payment orders has to be taken into account. Along with the volatile arrival rates of incoming orders there is a trade-off between idle times or delayed execution respectively.

As there are different providers for banking transaction services, an excess capacity market can be established using the corresponding technologies and standards described above. Then excess capacity can be traded among all transaction banks connected, i. e. orders the financial service provider submits into the IT system of its transaction bank are not executed by its respective in-house operating unit but with capacity of the excess capacity market if this execution path results in lower processing costs.

Characteristics of the Payment Transactions Process

The characteristics of the payment transaction process necessary to apply our model are identified as follows: Orders are accepted every bank working day between 7 a.m. and 10:00 p.m. Analyzing historical data reveals different peaks concerning the arrival rate of the incoming orders depending on exogenous

factors like billing cycles of the central bank or closing time. Dividing the 15 hours of order acceptance in seven time-frames, the arrival rate within each time-frame can be approximated by an exponential distribution with different means as summarized in Table 1.

Table 1. Arrival Rates Within a Bank Working Day (Mean Number of Orders per Minute)

7:00 a.m. – 8:30 a.m.	8:30 a.m. – 2:00 p.m.	2:00 p.m. – 3:00 p.m.	3:00 p.m. – 6:30 p.m.	6:30 p.m. – 8:00 p.m.	8:00 p.m. – 9:30 p.m.	9:30 p.m. – 10:00 p.m.
60	3	30	20	4	50	3

The processing of an order requires one unit of capacity for about 4:00 minutes in average, not dependent of the execution path. In this special case, idle capacity cannot be used to accelerate the execution of orders as only one employee can work on one order. Cost accounting reveals that one unit of capacity within the in-house operating unit causes fixed costs amounting to EUR 240 a bank working day.

The service level agreement between the financial service provider and the transaction bank consists of two deadlines: First, each order has to be processed within 12 minutes after arrival. For each minute an order exceeds this time frame, a compensation amounting to EUR 0,033 per minute is due. Second, there is a final processing deadline for each bank working day: All incoming orders have to be processed until 12:00 midnight. For each order which is not processed within this deadline the compensation payment rises to a penalty of EUR 51.

In the financial services industry it is necessary to execute payment orders in time, e. g. to meet external deadlines set by the central bank and to avoid losses of interest. Especially for legal and reputational reasons it is necessary that no order is left unexecuted. With the penalty connected to the final processing deadline of a bank working day the financial service provider places a high incentive to execute all incoming orders within a bank working day. Compared with the revenues earned with processing an order, this penalty for the final processing deadline is prohibitive.

For sake of simplicity, the variable costs for one order routed to the excess capacity market are fixed to EUR 1.96 within the simulation of this application scenario. Based on historical data provided from the external cloud service providers forming the excess capacity market the waiting time for excess capacity can be approximated. During a bank working day three time frames are identified. Each time frame shows different waiting times for free capacity which can be approximated by a normal distribution as outlined in Table 2. Orders routed to the excess capacity market have to wait according to the time frame valid at the time the order is routed to the excess capacity market. With the characteristics the discrete event simulation now can be set up.

Table 2. Distribution Parameters of the Waiting Time in Queue in Front of the Excess Capacity Market (Mean and Standard Deviation in Minutes)

7:00 a.m. – 12:00 noon	12:00 noon – 6:00 p.m.	after 6:00 p.m.
μ = 16:40; σ = 4:00	μ = 12:00; σ = 2:10	μ = 10:00; σ = 4:00

Discrete Event Simulation Set-Up

To examine the effect of exchanging excess capacity, the simulation has to run with and without the excess capacity market to determine the optimal amount of capacity to be allocated to the in-house operating unit. For both cases we proceed as follows: We execute multiple *simulation experiments* with increasing integer values for the capacity of the in-house operating unit. Each experiment consists of 1,000 *simulation runs*. For each run the total processing costs are determined. Starting the experiments with one unit of in-house capacity, we increase the value by one unit before the next experiment is started. This is done until the results of an experiment show that no waiting costs occur in front of the in-house operating unit for all runs. From this it follows, that a further increase of capacity does not have any positive effect of the total processing costs. Finally, comparing the average total processing costs for each experiment and choosing the one with the lowest costs then leads to the optimal in-house capacity.

With regard to the simulation time it is convenient that all bank working days of our application scenario are independently of each other (e. g. no unexecuted orders left due to the processing deadline at 12:00 midnight) and the relevant events which determine the optimal in-house capacity are recurrent each bank working day. Therefore it is sufficient to determine the optimal in-house capacity for a single day.

For each simulation run incoming orders are generated randomly following their statistical distributions. Whenever a new time frame is reached, the arrival rate is adapted. Concerning the availability of excess capacity, a random value is generated from the corresponding statistical distribution at the beginning of each time frame outlined in table 2. This value applies as the approximated waiting time for free capacity for the whole time-frame. Repeating a simulation run 1,000 times the risks connected to the waiting times for excess capacity are considered when using the results to determine the optimal in-house capacity.

Furthermore the routing algorithm is implemented determining the current processing costs of both paths each time an order arrives. Then it routes the order to the path with lower costs. Thereby the processing costs of the in-house execution result from the service level agreement with the financial service provider only. There are no variable costs and all fixed costs are sunk costs which must not be taken into account. From the service level agreement costs can occur in two different ways: If an order cannot be processed ahead of the final processing deadline, the penalty has to be considered within the processing costs. Otherwise, if the agreed processing time per order is exceeded costs per minute are charged. For the external execution the processing costs consist of the variable cost per order and the costs resulting from the service level agreement determined analogous.

Simulation Results

Figures 4 and 5 on the following page show the average costs and the average number of orders executed internally and externally depending on the capacity of the in-house operating unit without and with the excess capacity market.

Figure 4 reveals the influence of the cost associated with the service level agreement on the total processing costs: Very small in-house capacity results in a high amount of unexecuted orders and the total processing costs are very high due to the corresponding penalties. With increasing capacity, more orders are executed during a bank working day ahead of the final processing deadline and the total processing costs decrease accordingly. Increasing capacity furthermore implies an interesting effect on the service level costs for orders exceeding the agreed processing time: First, the increasing amount of orders which can be executed during a bank working day is accompanied by long waiting times in front of the in-house operating unit. That is why the corresponding costs rise up to more than EUR 100,000 maximum. Reaching a certain point, additional capacity not only executes more orders but also reduces the waiting times for these executed orders in the queue in front of the in-house operating unit. This explains the decreasing waiting costs with a capacity of more than about 42 units. These costs along with the fixed costs of capacity shape the total operating costs to a convex graph with a global minimum.

Figure 5 shows a completely different shaped graph of the total processing costs when the external execution path is added. The most striking change is the reduced range of variation of the total processing costs depending on the in-house capacity. Adding an external execution path therefore reduces the risk of allocating an inappropriate amount of capacity to the in-house operating unit. As the volatile arrival rates which determine the amount of capacity needed are estimated based on historical data but may change in the future this is an outstanding improvement concerning the capacity optimization problem for the cloud service provider. The reduced range of variation is due to the fact the excess capacity market allows the execution of orders which are left unexecuted in the other case. Without this external execution path, assigning too little capacity to the in-house operating unit inevitable leads to unexecuted orders. Excess capacity in contrast allows the execution although it is associated with variable execution costs and risky waiting times. The high penalty ensures the routing algorithm chooses the external execution path. Therefore the peak of the total processing costs occurring in the case without the external execution path with a capacity lower than about 75 units (as observable in Figure 4) is cut off.

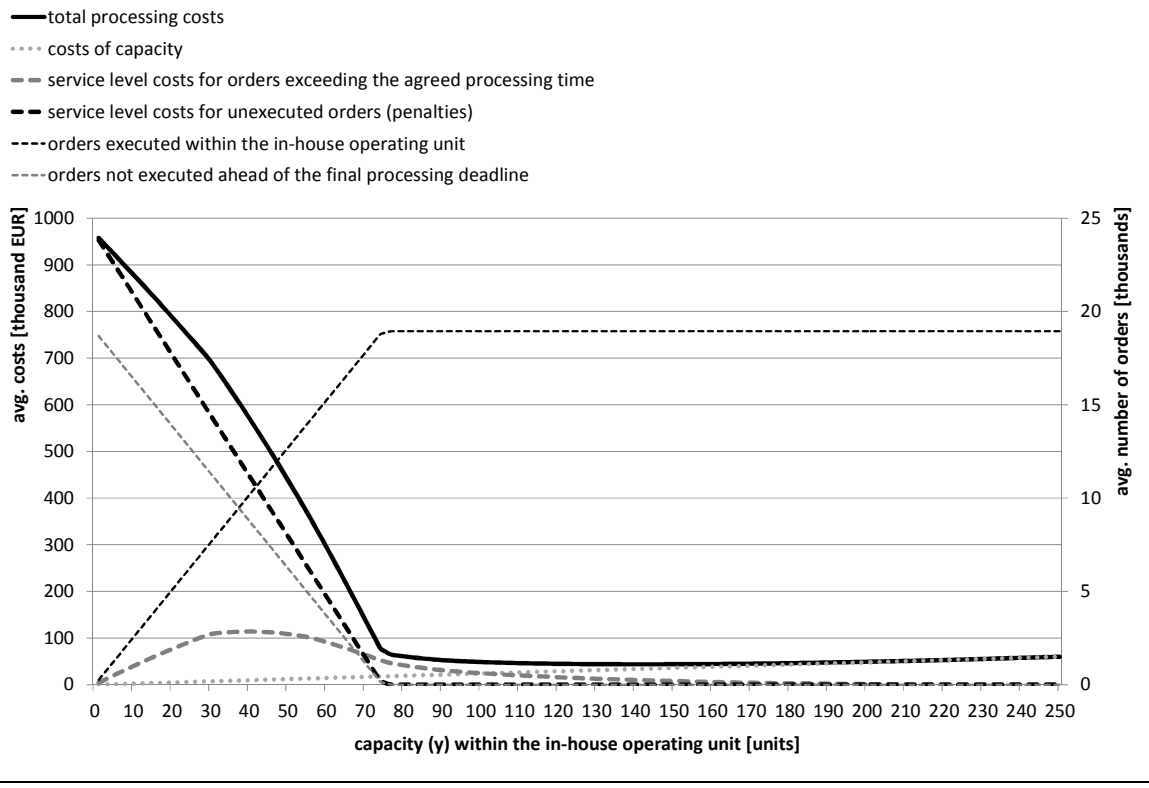


Figure 4. Simulation Results Without Excess Capacity Market

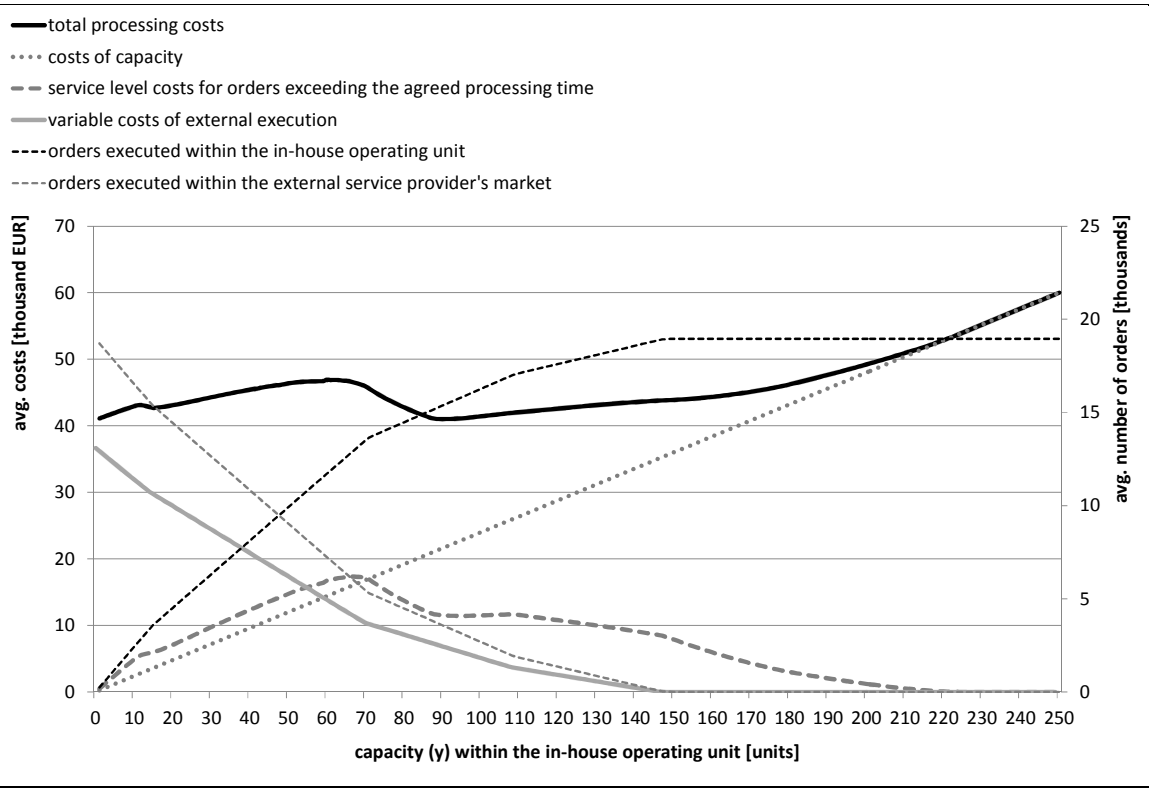


Figure 5. Simulation Results With Excess Capacity Market

But there is another reason for the different shape of the total processing costs: With excess capacity not only more orders are executed ahead of the final processing deadline. Moreover it reduces the waiting times in the queue in front of the in-house operating unit and for this reason the waiting costs. Orders arriving in peak times (e. g. the early morning with 60 orders per minute) can be routed to the excess capacity market and do not have to queue up in front of the in-house operating unit where the waiting times increase. This effect becomes clear if the graphs of the waiting costs in Figure 4 and 5 are compared. Without excess capacity waiting costs rises up to over EUR 100,000 maximum; with excess capacity the maximum is reached at about EUR 18,000.

Furthermore it may be mentioned that the waiting costs in front of the in-house operating unit are shaped similar to scenario one. Again, with increasing capacity more orders are executed internally and the waiting costs rise until the point additional capacity ensures reduced waiting times in front of the in-house operating unit.

Economic and Environmental Effect of Exchanging Excess Capacity

Relying on these results we are now able to examine the effect of exchanging excess capacity on economic efficiency and environmental sustainability as formulated in our research questions. Therefore we compare the simulation results without and with the excess capacity market.

Identifying the capacity level with lowest average total processing costs leads to the optimal amount of capacity, the cloud service provider should allocate to the in-house operating unit. Table 3 shows the corresponding results in an overview.

Table 3. Optimal Capacity of the In-house Operating Unit

excess capacity market	without	with	Δ
optimal capacity [units]	142	90	52
total processing costs [EUR]	43,747.96	41,011.92	2,736.04

The figures show that the use of the excess capacity market results in a sharp drop of capacity to be allocated in-house. The necessary capacity to reach the economic optimum is reduced by 36.6%. At the same time, the corresponding total processing costs can be reduced as well by 6.25%.

From the *economic point of view*, there are two main advantages in using excess capacity: Obviously, the total processing costs can be reduced. But also the variation of total processing costs depending on the in-house capacity is reduced (see Figure 5), thus decreasing the risk of allocating an inappropriate amount of capacity to the in-house operating unit. These advantages are gained although the use of excess capacity carries risk as there are no contractual agreements regarding the availability of excess capacity.

From the *environmental point of view* the sharp drop of capacity which has to be allocated to the in-house operating unit to achieve an optimal staffing has to be emphasized. The co-operation of ex ante planned capacity within the in-house operating unit and excess capacity has a substantial effect on the amount of in-house capacity required. At the same time, excess capacity of external cloud service providers formerly remained idle is now utilized via the market.

Using the excess capacity market a service provider has two possibilities to optimize its capacity (IT resources as well as employees) with regard to volatile demand: First, an *internal optimization* of capacity can take place. This is a known way of reducing the corresponding costs as exemplarily mentioned in the introduction speaking about the utilization of data centers and refers to the simulation executed without the excess capacity market. Second, implementing technologies and standards connected to architectural concepts like service-orientation and web-services enable the exchange of excess capacity through a quick and frictionless integration between IT systems and the accompanying provision of information. This induces the additional opportunity of an *external optimization*. Together, idle capacity can be reduced as the total amount of capacity a service provider assigns to a certain service is reduced and existing excess capacities in times with low demand are utilized via the excess capacity market.

Limitations and Directions for Further Research and Validation

Given the research gap identified in the related work section, the presented model is a first attempt to examine the economic and environmental effects of an IT-enabled *excess capacity* market for *services*. For a first step towards a deeper understanding, we focused on the perspective of a single cloud service provider with a capacity optimization problem using excess capacity from the market. Thereby we relied on the simplifying assumption of an *exogenous market*. This means first, the amount of excess capacity available on the market is not affected by the actual demand of the cloud service provider or any other user of the market. And second, no dependencies are considered between peak times the cloud service provider or the market players are facing, which would have effect on the match of supply and demand on the market. This assumption of an exogenous market was necessary as corresponding functional relationships or knowledge about strategic behavior which could have been considered at this point of the model are not yet examined in detail (especially not for the new and exemplary application scenario) and this examination was not within the scope of this paper.

Considering the effects of an *endogenous market* within the model would have effects on the availability of excess capacity influencing the optimization result for the cloud service provider. In particular, this is the case if all external cloud service providers which form the market would apply the same unbalanced strategy (keeping very little/much in-house capacity relying mainly on excess capacity supply/demand) without carefully observing the availability of excess capacity on the whole market. Then the positive economic and environmental effects identified in this paper could be undermined in two ways: Either, this behavior could result in an excessive supply of capacity facing no adequate demand. Or the availability of excess capacity is very limited and as a result, the actual demand could not be satisfied.

Although the application scenario presented in the previous section shows reasonable results based on an *exogenous excess capacity market*, this short discussion points out the main limitation of our model and may serve as a starting point for further analytical and empirical research. Both should thereby focus the interdependencies between the excess capacity market and the strategic behavior of the market players:

From an *analytical point of view*, modeling an endogenous market is a promising subsequent step. Then the changing availability of excess capacity due to the actual demand as well as all players of the excess capacity market along with their corresponding behavior can considered. And the overall benefits due to the optimization efforts of all cloud service providers can be examined and possible market equilibrium may be determined. Furthermore, the effects concerning different characteristics of excess capacity can be examined, e. g. if the market players are faced with parallel or opposed peak times, as this is the case, when cloud service providers of different industries offer their excess capacity on the same market.

From an *empirical point of view* two main starting points are given: First, with detailed case studies relying on field data, the applicability of the model can be validated and an expanded set of effects in different application scenarios, e. g. within different industries, can be determined. Furthermore, with detailed field experiments the results found by analyzing the analytical model could be substantiated and the model robustness could be validated. This seems to be valuable especially with regard to the most important input parameters like the availability of excess capacity or the peak-times of order-arrival. Second, to address the lack of knowledge concerning interdependencies between the strategy of a single player and an endogenous excess capacity market appropriate field studies are necessary. These studies can provide deeper insights in the behavior of market players in their particular environment. The gathered data then can be used not only for qualitative insights about possible strategies an excess capacity market for services offers to the different players but furthermore builds the basis to identify quantitative functional interdependencies for further analytical work as outlined above.

Another link for *subsequent empirical research* concerns the data the ex ante planning of capacity relies on. Arrival rates of incoming orders and execution times are traceable or can be derived from contractual agreements. But it may not be reasonable to simply project it into the future. Rather, the exploration of concrete dependencies based on empirical data may lead to better forecasts. The same applies to data necessary to estimate the availability of excess capacity on the excess capacity market. Furthermore, in addition to comprehensive empirical investigations to gather profound knowledge of all necessary characteristics of the market, a careful empirical approach to determine the amount of capacity required for economic efficiency, e. g. by field studies might be useful in addition to optimization results found within the analytical model.

Summary

Applying the design science research approach, we build a mathematical model to examine the economic and environmental effects of an IT-enabled market which allows the exchange of excess capacity for services. Thereby we focus on a single cloud service provider which offers a service to its customers and uses excess capacity in order to support its in-house operating unit and to solve the capacity optimization problem resulting from volatile demand.

With our model we address a wide range of possible application scenarios found in nearly all companies and industries. Referring to the business process perspective of an enterprise, the application of the model requires a digitalized process where orders or resources can be exchanged via electronic networks and at the same time a standardized process for which corresponding services are offered by external service providers. Currently, this applies mainly to supporting or back-office processes like payroll accounting, marketing campaign management or applicant management. However, it can be expected, that process standardization due to different sourcing decisions leads to wider application possibilities. Especially in industries like the financial services sector with digitalized products, high standardization and a fragmented supply chain with numerous service providers for different services, even core processes comply with these requirements.

Hence, we choose the financial services sector for a computer-aided simulation to analyze and evaluate the mathematical model. In doing so, we found reasonable benefits of exchanging excess capacity concerning the economic and the environmental perspective. Being the first quantitative approach, the presented model thereby builds an analytical basis for both, empirical validation and testing of the cloud service relationship regarded as well as further analytical research concerning the interdependencies and relationships between the market and its individual players.

References

Adenso-Diaz, B., Gonzalez-Torre, P., and Garcia, V. 2002. „A capacity management model in service industries," *International Journal of Service Management* (13:3), pp. 286-302.

Allon, G., and Federgruen, A. 2006. "Outsourcing Service Processes to a Common Service Provider under Price and Time Competition," *Working Paper*, Kellogg School of Management, Northwestern University, Evanston, IL, pp. 1-50.

Aksin, Z., de Vericourt, F., and Karaesmen, F. 2008. "Call Center Outsourcing Contract Analysis and Choice," *Management Science* (54:2), pp. 354-368.

Anandasivam, A., and Premm, M. 2009. "Bid Price Control and Dynamic Pricing in Clouds," *Proceedings of the 17th European Conference on Information Systems*, ECIS, Verona, Italy, pp. 1-14.

Baliga, J., Ayre, R. W. A., Hinton, K., and Tucker, R. S. 2010. "Green Cloud Computing: Balancing Energy in Processing, Storage, and Transport," *Proceedings of the IEEE* (99:1), pp. 149-167.

Bassamboo, A., Ramandeep, S. R., and van Mieghem, J. A. 2010a. "Optimal Flexibility Configurations in Newsvendor Networks: Going Beyond Chaining and Pairing," *Management Science* (56:8), pp. 1285-1303.

Bassamboo, A., Randhawa, R. S., and Zeevi, A. 2010b. "Capacity Sizing Under Parameter Uncertainty: Safety Staffing Principles Revisited," *Management Science* (56:10), pp. 1668-1686.

Bohrer, P., Elnozahy, E. N., Keller, T., Kistler, M., Lefurgy, C., McDowell, C., and Rajamony, R. 2002. "The case for power management in web servers," *Power Aware Computing,* Graybill, R., and Melhem, R. (eds.), Norwell: Kluwer, pp. 261-289.

Braunwarth, K., and Ullrich, C. 2010. "Valuating Business Process Flexibility achieved through an alternative Execution Path," *Proceedings of the 18th European Conference on Information Systems*, ECIS, Pretoria, South Africa, pp. 1-13.

Cachon, G. P., and Harker, P. T. 2002. "Competition and Outsourcing with Scale Economies," *Management Science* (48:10), pp. 1314-1334.

Chen, A. J., Watson, R. T., Boudreau, M., and Karahanna, E. 2009. "Organizational Adoption of Green IS & IT: An Institutional Perspective," *Proceedings of the 30th International Conference on Information Systems*, ICIS, Phoenix, AZ, USA, pp. 1-18.

Chen, L., and Nunez, M. 2010. "Business Process Integration of Multiple Customer Order Review Systems," *IEEE Transactions on Engineering Management* (57:3), pp. 502-512.

Chesbrough, H., and Spohrer, J. 2006. "A Research Manifesto for Service Science," *Communications of the ACM* (49:7), pp. 35-40.
Cook, G. 2012. "How clean is your cloud," http://www.greenpeace.de/fileadmin/gpd/user_upload/themen/klima/HowCleanisYourCloud_final.pdf.
Dong, L., and Durbin, E. 2005. "Markets for Surplus Components with a Strategic Supplier," *Naval Research Logistics* (52), pp. 734-753.
Dorsch, C., and Häckel, B. 2012. "Integrating Business Partners On Demand: The Effect on Capacity Planning for Cost Driven Support Processes," *Proceedings of the 45th Hawaii International Conference on System Science,* HICSS, Maui, Hawaii, pp. 4796-4805.
Gans, N., and Zhou, Y.-P. 2007. "Call-Routing Schemes for Call-Center Outsourcing," *Manufacturing & Service Operations Management* (9:1), pp. 30-50.
Grefen, P., Ludwig, H., Dan, A., and Angelov, S. 2006. "An Analysis of Web Services Support for Dynamic Business Process Outsourcing," *Information and Software Technology* (48), pp. 1115-1134.
Gross, D., Shortle, J. F., Thompson, J. M., and Harris, C. M. 2008. "Fundamentals of Queuing Theory," Hoboken, NJ: Wiley.
Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. "Design Science in Information Systems Research," *MIS Quarterly* (28:1), pp. 75-105.
Hlupic, V., and Robinson, S. 1998. "Business process modeling and analysis using discrete-event simulation," *Proceedings of the 1998 Winter Simulation Conference*, Washington DC, USA, pp. 1363-1369.
Lee, H., and Whang, S. 2002. "The Impact of a Secondary Market on the Supply Chain," *Management Science* (48:6), pp. 719-731.
Liu, J., Zhao, F., Liu, X., and He, W. 2009. "Challenges towards elastic power management in internet data centers," *Proceedings of the IEEE International Conference on Distributed Computing Systems,* Los Alamitos, CA, USA, pp. 65–72.
Liu, T. 2010. "Revenue Management model for on-demand IT services," *European Journal of Operations Research* (20:7), pp. 401-408.
McKay, J., Marshall, P., and Hirschheim, R. 2012. "The design construct information systems design science," *Journal of Information Technology* (27), pp. 125-139.
Melville, N. P. 2010. "Information Systems Innovation for Environmental Sustainability," *MIS Quarterly* (34:1), pp. 1-21.
Meredith, J. R., Raturi, A., Amoako-Gyampah, K., and Kaplan, B. 1989. "Alternative Research Paradigms in Operations," *Journal of Operations Management* (4), pp. 297-326.
Moitra, D., and Ganesh, J. 2005. "Web Services and Flexible Business Processes: Towards the Adaptive Enterprise," *Information & Management* (42), pp. 921-933.
Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. 2008. "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems* (3), pp. 45-77.
Rai, A., and Sambamurthy, V. 2006. "Editorial Notes – The Growth of Interest in Service Management: Opportunities for Information Systems Scholars," *Information Systems Research* (17:4) pp. 327-331.
Ren, Z. J., and Zhou, Y.-P. 2008. "Call Center Outsourcing: Coordinating Staffing Level and Service Quality," *Management Science* (54:2), pp. 369-383.
Singh T., and Vara, P. K. 2009. "Smart metering the clouds," *Proceedings IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises*, Groningen, The Netherlands, pp. 66–71.
The Climate Group 2008. "Smart2020: Enabling the Low Carbon Economy in the Information Age," http://www.smart2020.org/_assets/files/02_Smart2020Report.pdf.
Vereecken, W., Deboosere, L., Colle, D., Vermeulen, B., Pickavet, M., Dhoedt, B., and Demeester, P. 2008. "Energy efficiency in telecommunication networks," *Proceedings 13th European Conference on Networks and Optical Communications*, Krems, Austria, pp. 44–51.
Vykoukal, J. 2010. "Grid Technology as Green IT Strategy? Empirical Results from the Financial Services Industry," *Proceedings of the 18th European Conference on Information Systems*, ECIS, Pretoria, South Africa, pp. 1-13.
Wacker, J. G. 1998. "A definition of theory: research guidelines for different theory-building research methods in operations management," *Journal of Operations Management* (4), pp. 361-385.

Watson, R., Aronson, J., Donnellan, B., and Desautels, P. 2009. "Energy + Information < Energy," *Proceedings of the 15th Americas Conference on Information Systems*, AMCIS, San Francisco, CA, USA.

Watson, R. T., Boudreau, M., and Chen, A. J. 2010. "Information Systems and Environmentally Sustainable Development: Energy Informatics and New Directions for the IS Community," *MIS Quarterly* (34:1), pp. 23-38.

Weinhardt, C., Anandasivam, A., Blau, B., Borissov, N., Meinl, T., Michalk, W., and Stößer, J. 2009. "Cloud Computing – A Classification, Business Models, and Research Direction," *Business & Information Systems Engineering* (5), pp. 391-399.

Wurman, P. 2001. "Dynamic Pricing in the Virtual Marketplace," *IEEE Internet Computing* (5:2), pp. 36-42.